Procesamiento del Lenguaje Natural, núm. 35 (2005), pp. 425-432 recibido 30-04-2005; aceptado 01-06-2005

Identifying Jargon in Texts

Jesús Llevadias Jané
Universitat Politècnica de Catalunya
Barcelona, Spain
jllevadias@crl.nmsu.edu

Stephen Helmreich
Computing Research Laboratory
New Mexico State University
shelmrei@crl.nmsu.edu

David Farwell
Computing Research Laboratory and
Universitat Politècnica de Catalunya
david@crl.nmsu.edu
david@lti.upc.es

Resumen: Este ponencia describe un proyecto actual sobre en el uso de jerga para identificar las creencias del autor de un texto. El enfoque central de la investigación es el uso de técnicas estadísticas para ayudar a identificar terminología que caracteriza la jerga de grupos que se comparten ciertas creencias sobre un tema. El tesis central es que un grupo que se comparten un rango de creencias tiende a desarrollar una manera especial para hablar de este tema, y en particular una que sea distinta de los que comparten creencias alternativas. Describimos un programa de investigación con tres etapas y las decisiones tomados a cada paso. Presentamos unos resultados preliminarios y los planes para la contiuación del proyecto.
Palabras clave: jerga, creencias del autor, clasificación de textos

Abstract: This paper describes on-going work on using jargon to identify author's beliefs. The main focus of research is to use statistical techniques to help identify particular lexical items that characterize the jargons of groups with particular beliefs. The main thesis was that groups that hold common beliefs tend also to develop common ways of speaking about the topic of those beliefs, ways that differ from those who hold differing beliefs about the same subject area. We describe a three-step program of research and the decisions and results reached at each step. Preliminary results are reported and continuing work described.
Keywords: jargon, author beliefs, text classification, machine learning

1 Introduction

This paper reports on the results of an experiment to determine whether or not it is possible to identify a jargon for social subgroups, in this case, groups that are associated with opposite sides of a polarized issue (in this case, abortion). This experiment is part of a research program (funded under NSF ITR # IIS-0313338) to determine the validity of a central hypothesis: that it is possible in general to use jargon to identify various attributes of the author of a text or, more specifically, the author's opinions or system of beliefs. This involves the subhypothesis that it is possible to identify the jargons of particular groups. The key to this research is the assumption that to identify the author's opinions or beliefs it is sufficient to identify the groups to which the author belongs by identifying the jargon or jargons the author uses in writing a text.

By "opinion," we mean roughly a position taken with respect to a single issue. By "system of beliefs," we mean a set of related beliefs held in common by a social group (often of a philosophical or foundational nature). By "jargon," we mean a set of vocabulary items, collocations, or formulaic constructions which are used among people holding similar opinions or sharing a common system of beliefs. Such jargons, if they exist, should reliably correlate with the texts produced by people who share those opinions or systems of belief, and in particular those texts which express such opinions or promote such a system of beliefs. Such jargons should be distinguishable from jargons associated with those holding contrasting opinions or having an alternative system of beliefs. For example, someone who favors free access to abortion might use the

ISSN: 1135-5948

word *fetus* in referring to a developing embryo inside its mother's womb, while those favoring limited access might use the term *unborn child*. Similarly, a Roman Catholics use the word *priest* to refer to a religious leader, Protestants generally use the words *minister* or *pastor*, while *rabbi*, and *imam* are used by Jews and Muslims respectively.

The question at issue here is whether use of this jargon extends to texts that are not specific to the domain relevant to the beliefs of the group. Do terms specific to certain religious communities, for example, get used by its members in texts that are not specifically about religious issues nor directed specifically at other community members? If so, use of such terms would provide a clue as to the community to which the author belongs. For instance, if an author refers to a generic religious leader as a "priest," that might indicate that the author is a member of the Roman Catholic community.

The research reported here deals with a first step in establishing this hypothesis, namely that it is possible to extract jargons that can reliably be used to identify two subgroups polarized around an issue, when examining texts that are related to that issue. The issues selected were abortion and gun control, which have highly polarized opposing points of view: pro-life or pro-choice and pro-right to bear arms or pro-control.

The results of this research show that indeed, jargon terms can be identified for these subgroups, and that these jargon terms can be used to reliably distinguish texts produced by these different groups.

The work reported on here consisted of several consecutive activities: corpus collection and preparation, proposing jargon candidates based on corpus analysis, selection a jargon, and finally using that jargon to reliably identify those authors that use that jargon. We report on each activity in turn and conclude with a discussion of future research.

2 Corpus collection and preparation

Two areas were chosen with well-defined subgroups: abortion and gun-control. (Only results from the abortion texts are currently available.) Corpus collection was done by web crawling. To be sure that our corpora were *representative samples* of the populations of interest, we looked only at websites specially connected with the chosen topics. We also determined for each website the viewpoint (supporting and opposing) for each topic. Consequently, we could be sure that our texts were representative of the different points of view. Finally, we gathered a corpus of about 1000 documents representing each subgroup for each topic.

Once the initial corpora were prepared, several decisions had to be made before statistical processing could begin. Because of the source of our corpora (the Internet), there was a lot of formatting and irrelevant content that needed to be filtered out. There were decisions to be made about standardization as well, such as whether or not to distinguish between capitalized and non-capitalized forms, e.g., between *National* as in *National Rifle Association* and *national*. For most of our experiments we worked at the word level and, as a result, part-of-speech was very helpful.

We prepared two versions of each corpus – one as is, the other lemmatized and categorized with part of speech. To do the tagging, we used the POST tagger (Weischadel *et al.,*1993), and the Penn Treebank parts of speech. Only five categories of words were kept (adjectives, adverbs, nouns, proper names and verbs). Table 1 summarizes information about the corpus.

Side	Documents	Words	Tagged Words
Pro-life	450	490,800	294,596
Pro-choice	412	384,390	227,731

Table 1. Corpus information

3 Looking for jargon candidates

Term or concept extraction systems give different results according to their search strategies. On one hand, systems that use statistical methods can treat large datasets but do not allow for a very rich interpretation of the results. On the other hand, systems using linguistic methods allow for the use of a finer semantics have a good deal of difficulty with larger datasets. According to this, we decided to create a system that merges these two approaches so as to better answer our particular needs. The sources we exploited were statistics, semantic relationships and expert knowledge. Figure 1 shows the how and when the different types of knowledge were applied.

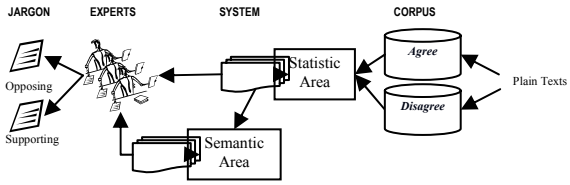

Figure 1. Flow of knowledge used to find the Jargon.

After obtaining the corpora, the first step was to examine them for distinctive terminology. Generally, to do this we used a program described in (Guthrie *et al.,* 1992). Given two corpora, it produces two lists, each of which contains words appearing frequently in one of the corpora but infrequently in the other.

However, we did not compare the two issue-based corpora directly. We wanted to know that the frequency difference between the two corpora was significant, and not just due to random variation in the standard frequency of the terms in question. So we compared each corpus with a large open domain English corpus derived from the Wall Street Journal (WSJ).

As a result each corpus list contained words that occurred more frequently in that corpus than in standard English. We created the word frequency lists based on the two forms of the corpora described in Section 2 – one based purely on word forms and other on part-of-speech-tagged citation forms.

In addition, we looked at significant bigrams, omitting from the text prepositions and articles, again using both versions of the corpus. This resulted in four sets of files of words and bigrams from each corpus that were used more frequently in that corpus than in standard English.

More specifically, for statistical processing we exploited a number of standard techniques such as word frequency counts and collocation identification methods. These collocation tools would include likelihood ratios and mutual information (Manning & Schütze, 1999).

The simplest method for trying to find jargon is counting. Our first experiment was to identify those words which occur in the corpus representing one of the opinion groups but do not occur at all in the corpus representing the other group. Although this resulted in a certain degree of positive results, for instance people that supported the right to choose had a marked tendency to use words like *physician, anti-choice* or *rights*, we detected a bias. We only identified terms with a significantly different distribution between the two points of view; but, this did not take into account the standard frequency of those words. What we wanted to know was that the frequency difference between the two corpora was significant, and not just due to random variation in the standard frequency of the words in question. So we compared each subgroup corpus with a large open domain English corpus from the WSJ in order to get some idea of which words were *significantly* different.

One difficulty initially is that high frequency terms for one opinion group for a particular issue could be just accidental. For example, word *site* occurs frequently in the Gun Control corpus and we saw that this was because *site* has a high frequency in both sub-corpora, so we expected that the word was associated with one particular group or the other just by chance. This can be couched in terms of Hypothesis Testing (HT). A null hypothesis, H_0, is formulated such that there is no association between terms and groups beyond chance. Compute the probability p that the event would occur if H_0 were true, and then reject H_0 if p is too low (the significance level was $p<0.05$) and retain H_o as possible otherwise. Maximum Likelihood was the first approach used to compute the HT because it has a clear intuitive interpretation. Given a list of words with their maximum likelihood ratio, we could reject or accept the null hypothesis for each. Second approach used was Mutual Information, and we used it as a measure of association between elements (how much one event tells us about the other).

For example, the two subcorpora for gun control correspond to people who support the right to own a gun on the one hand and to those who favor some form of gun control on the other. By computing mutual information we learn that the amount of information we have about the occurrence of a term, for example, *freedom*, increases by 0.74493533 if *freedom* occurs in the subcorpus of the group against the gun control. In other words, we can be much more certain that *freedom* will occur if the texts are by people against the gun control. Unfortunately, this measure of "increased information" is in some cases not a useful measure given our interests.

Consider the example in Table 2 of counts of word *Home* tagged as a proper noun (PN).

Abortion	Agree	Disagree
HomePN	402	37

Table 2. – Number of occurrences of *HomePN* in Abortion Corpora

The reason that *HomePN* appears frequently in the pro-choice corpus was that the websites we accessed for that group contained a link to the *Home Page* for each web page, article or discussion mentioned. Consequently, *HomePN* for the abortion corpus was very frequent for the pro-choice group and very infrequent for the pro-life group. Mutual Information, of course, gave a higher value to *HomePN* and that higher value reflects the fact that *HomePN* results in a larger decrease in uncertainty with respect to the topic. Nevertheless, as the example shows, a decrease in uncertainly does not necessarily correspond to what we consider to be good jargon term.

In regard to semantics, one of the more important clues to possible membership in some jargon is the atypical use of terminology to refer to particular objects or concepts. So, for instance, whereas a member of ETA might refer to a particular person as a *freedom fighter* or *soldier*, many others might refer to that same person as a *terrorist*. To identify such usages, we expect to use WordNet (Fellbaum, C. (ed.). 1998) and Omega (http://omega.isi.edu) semantic resources. Focus is on terms in common usage, but which have a special sense or connotation. Uses of these words with these special senses in other texts are clues to the use of that jargon in the text.

In addition to variation in the selection of lexical items from synonym sets, we also expect that there are variations in the directness or indirectness of reference by way of hypernyms. For instance, in the abortion sub corpora we frequently found the use of *human* to refer to a *fetus*. Again, to the degree possible, semantically-based word list resources such as those mentioned above were used to identify such lexical relationships.

A major goal was to isolate lexical items in the two corpora that refer to the same event, object, or relation, but by different terms. The hypothesis is that these are especially good terms by which to identify the beliefs of the author.

To this end, a tool has been created that finds terms in each of two corpora that are in the same semantic field. For WordNet these are terms inside the same synset and for Omega, they are terms connected to the same ontological node.

Using WordNet, for example, we found a parent relationship between *abortionist*, which is common in the pro-life subcorpus, and *physician*, which is common in the pro-choice subcorpus, both of which share the same mother, i.e., *doctor*. On the other hand, for the gun control corpus, Omega gave us a sister relationship between *murder* and *homicide*, both of which are under the concept *slaying*. These pairs, then are particularly indicative of an author's opinions.

4 Jargon selection

Having produced lists of potential jargon terms for each of the four opinion groups (each containing over 1000 lexical items and two-word collocations), we then examined each list by hand and selected a much smaller set (approximately 100) to serve as the actual set of jargon words. In making the selection, we avoided words that were related to content that characterized one set but not the other, focussing instead on words related to content common to both corpora but reflecting different connotations. For example, pro-choice texts often focussed on the bombing of abortion clinics and so the terms *bomb* and *bombing* appeared more frequently in these texts. However, in terms of characterizing a pro-choice jargon these terms have limited value.

On the other hand, pro-life texts referred frequently to *crisis pregnancies* while pro-choice texts referred to *unintended pregnancies*. These terms we believe indicated a difference in conceptualization of these pregnancies and thus would be useful in distinguishing pro-life from pro-choice texts and in identifying abortion-related jargon terms in texts perhaps not specifically about abortion. Other terms selected included clearly reciprocal designations of the other's position (*pro-choice* vs *pro-abortion*; *pro-life* vs *anti-choice*). Some terms were included that carried quite negative connotations – *abortionist* versus *doctor*, but some did not – *doctor* versus *physician*.

5 Using jargon to classify texts

The first step in applying jargon to the classification of texts is to find an appropriate data representation. This is an art in itself, and

usually depends on the categorization method used. The representations chosen were directly based on the jargon terms for abortion or gun control opinion groups. Given an opinion group with its jargon word list, each document, *j*, that makes up the test corpus is represented as a vector of *k* integers $\vec{x} = (s_{1,j},...,s_{k,j})$, where *k* is the number of jargon words we have for this the opinion group and $s_{i,j}$ is computed as a binary or weight representation.

The first is the simplest representation, consisting of a matrix of documents where columns are jargon words and rows are the documents that form the corpus. If cell $s_{i,j}$ is equal to one means that word *i* appears one or more times in document *j*, if it is equal to zero, it means the term does not appear in the document.

For the second, a cell is equal to

$$s_{i,j} = 10 * \frac{1+\log(tf_{i,j})}{1+\log(l_j)}$$

For this representation $tf_{i,j}$ is the number of occurrences of term *i* in document *j* and l_j is the length (in number of words) of document *j*. The score $s_{i,j}$ is set to 0 for non-occurrences of the term. Finally, we add an additional attribute or type to both representations that tells us the classification of each document as pro-choice or pro-life..

The second step in applying jargon to the classification of texts is to find a class model. We used the University of Waikato's WEKA toolkit, http://www.cs.waikato.ac.nz/ml/weka/ to generate this. In all, we experimented with three different types of algorithms: OneR, ID3, and C4.5.

The Rule Based Model uses the OneR algorithm for implementation. It produces a very simple decision based on only one attribute However, it is powerful enough to detect possible biases that could be resolved for a posteriori analysis with more complex models. For example, we ran the model for the abortion issue and we noted an incredibly high performance for this simple model. Using the noun *priest*, we got a performance of 94.0698% for correctly classified documents. The reason was that one of the largest websites where we got the data for the pro-life group was *www.priestforlife.com*. They put the name of their website in most of the pages as a header. Consequently, *priest* was an easy word to discriminate between the two valid positions. After this, we decided to eliminate *priest* as a possible candidate because it introduced too much noise for our analysis.

We then looked at two types of decision tree models, ID3 and C4.5. We adopted them for two main reasons. On the one hand, we wanted an analysis tool to test the main hypothesis, and, on the other, we wanted to determine the role that each jargon word plays in the classification process. Trees permit a mental fit; but, in order to be comprehensible, the rules induced from the tree need to be as short as possible. However, over fitting with ID3 can lead to long rules. In order to induce shorter rules, it is usually necessary to relax the requirement that the induced rules be consistent with all the training data as C4.5 does.

6 *Results and discussion*

We used four versions of the jargon terms, crosscutting two variables: the use of word form vs the use of citation form (and part of speech information); the use of occurrence data vs use of frequency of occurrence data. Thus the following four datasets were used: (1) word form / occurrence data; (2) citation form / occurrence data; (3) word form / frequency of occurrence data; (4) citation form / frequency of occurrence data.

The OneR algorithm selects only one feature to make the decision. It produced the lowest accuracy results as shown in Table 3. This was based on a 2/3-1/3 split in the corpus: 2/3 for training, 1/3 for testing.

Data Set	Training Set (%)	Testing Set (%)	Appendix
1	92.6914	95.2381	OneR.1
2	93.6342	94.5578	OneR.2
3	88.9017	89.1525	OneR.3
4	93.6342	95.2381	OneR.4

Table 3. Results of OneR algorithm

The ID3 algorithm produced very good results, but the trees are quite deep and dense, and thus a bit over trained. In addition, the input data must be whole numbers, thus excluding datasets 3 and 4 in which the frequency of occurrence is weighted by the length of the document, resulting in non-integer data. The results are shown in Table 4.

Data Set	Testing Set (%)	Appendix
1	97.2789	ID3.1
2	97.9592	ID3.2

Table 4. Results of the ID3 algorithm

The C4.5 algorithm produced the best results overall as shown in Table 5.

Data Set	Testing Set (%)	Appendix
1	99.3197	C4.5.1
2	96.9388	C4.5.2
3	94.9153	C4.5.3
4	97.2789	C4.5.4

Table 5. Results of the C4.5 algorithm

The results of the three experiments together are summarized in Figure 2.

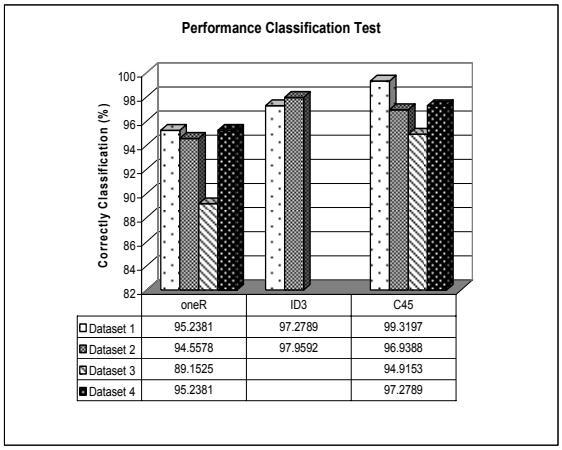

Figure 2. Summary of classifier results

We also performed a 10-fold cross-validation study that showed that the C4.5 algorithm produced the best overall results. Each run represents a 10-fold cross validation run, dividing the corpus into 10 parts, training on nine and testing on the tenth. The results are in Table
6.

Data Set	Number of Runs	OneR	ID3	C4.5
1	1	92.6914	95.7077	98.1439
	2	92.6914	96.4037	97.4478
	3	92.6914	95.8237	97.2158
	4	92.6914	95.7077	97.3318
	5	92.6914	96.5197	97.6798
	6	92.6914	95.9397	97.6798
	7	92.6914	95.9397	97.3318
	8	92.6914	96.5197	97.3318
	9	92.6914	96.1717	97.5638
	10	92.6914	95.9397	97.9118
	Mean	92.6914	96.0673	97.56381
2	1	93.6343	96.0648	96.5278
	2	93.6343	96.9907	96.6435
	3	93.6343	96.9907	96.875
	4	93.6343	97.1065	97.1065
	5	93.6343	96.7593	96.6435
	6	93.6343	97.6852	96.9907
	7	93.6343	96.0648	96.9907
	8	93.6343	97.2222	97.2222
	9	93.6343	97.1065	96.6435
	10	93.6343	97.2222	96.6435
	Mean	93.6343	96.9213	96.82869
3	1	88.9017		96.185
	2	88.9017		96.185
	3	88.9017		96.0694
	4	88.9017		95.1445
	5	88.9017		95.4913
	6	88.9017		95.4913
	7	88.9017		95.9538
	8	88.9017		96.185
	9	88.9017		96.185
	10	88.9017		95.1445
	Mean	88.9017		95.80348
4	1	93.6343		97.4537
	2	93.6343		97.5694
	3	93.6343		98.0324
	4	93.6343		98.1481
	5	93.6343		97.9167
	6	93.6343		97.5694
	7	93.6343		97.5694
	8	93.6343		97.6852
	9	93.6343		97.8009
	10	93.6343		97.8009
	Mean	93.6343		97.75461

Table 6: Results of 10-fold validation.

The results of the cross-validation study are summarized in Figure 3.

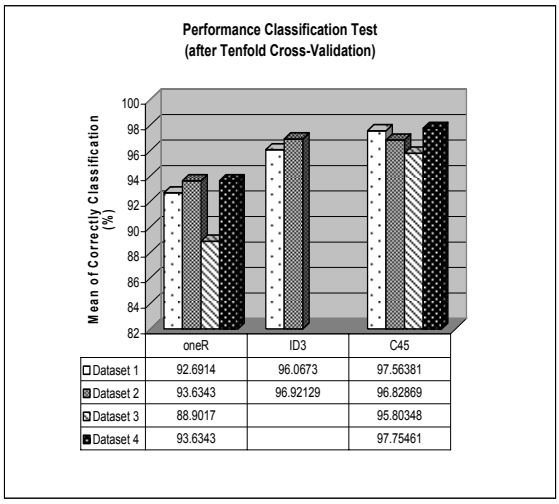

Figure 3. Table of Mean results

Finally, we applied a T-test to evaluate the confidence in the hypothesis. Here, if the T-test value is greater than Z value (T-test > Z), or if the T-test value is smaller than -Z (T-test < -Z), the hypothesis cannot be rejected. This is the case for the T-tests below (Table 7).

DATASETS WITHOUT WEIGHTS				
Data Set	Differences	ID3 vs OneR	C4.5 vs OneR	C4.5 vs ID3
1	D1	3.0163	5.4525	2.4362
	D2	3.7123	4.7564	1.0441
	D3	3.1323	4.5244	1.3921
	D4	3.0163	4.6404	1.6241
	D5	3.8283	4.9884	1.1601
	D6	3.2483	4.9884	1.7401
	D7	3.2483	4.6404	1.3921
	D8	3.8283	4.6404	0.8121
	D9	3.4803	4.8724	1.3921
	D10	3.2483	5.2204	1.9721
	Std Dev	0.31626402	0.29449845	0.47185053
	T-Test	***33.7551296***	***52.3191668***	***10.0294051***
2	D1	2.4305	2.8935	0.463
	D2	3.3564	3.0092	-0.3472
	D3	3.3564	3.2407	-0.1157
	D4	3.4722	3.4722	0
	D5	3.125	3.0092	-0.1158
	D6	4.0509	3.3564	-0.6945
	D7	2.4305	3.3564	0.9259
	D8	3.5879	3.5879	0
	D9	3.4722	3.0092	-0.463
	D10	3.5879	3.0092	-0.5787
	Std Dev	0.50949693	0.2390685	0.49044754
	T-Test	***20.4012514***	***42.2537806***	***-0.5970606***

DATSETS WITH WEIGHT		
Dataset	Differences	C4.5 vs OneR
3	D1	7.2833
	D2	7.2833
	D3	7.1677
	D4	6.2428
	D5	6.5896
	D6	6.5896
	D7	7.0521
	D8	7.2833
	D9	7.2833
	D10	6.2428
	Std Dev	0.43956564
	T-Test	***49.6520716***
4	D1	3.8194
	D2	3.9351
	D3	4.3981
	D4	4.5138
	D5	4.2824
	D6	3.9351
	D7	3.9351
	D8	4.0509
	D9	4.1666
	D10	4.1666
	Std Dev	0.22628231
	T-Test	***57.5810105***

Table 7. T-Test results

The results show that all of the comparisons are valid because they are all outside of the Z interval (*[-1.83,1.83]*). Consequently, none can be rejected. The OneR algorithm provided a performance benchmark – but on weighted Datasets 3 and 4, C4.5 was much better. For Datasets 1 and 2, the C4.5 algorithm is better than ID3 but not by much. However, the decision tree generated for the C4.5 algorithm is smaller than the tree generated for the ID3.

Although these results are still preliminary, we believe that they demonstrate an initial step in the validation of the central hypothesis, namely, that it is possible to determine an author's beliefs on the basis of the jargon the author uses. We believe these results justify additional research in this area.

7 Future work

Initial ongoing research includes completion of the work on the gun-control corpus. A

second phase will examine corpora from multiple social subgroups organized around a larger topic of concern, such as religion (e.g., Buddhism, Christianity, Hinduism, Islam, Judaism). We also plan to test the jargons on texts that are not directly about the issue or concern in specific. Both of these steps will involve categorization into multiple boxes.

We also intend to further investigate the possibility of semi-automatic methods for jargon selection. In particular, we will look at other lexical semantic ontological resources, such as Omega or OntoSem (Mahesh and Nirenburg, 1995) or Dekang Lin's method of determining semantic equivalents (Lin *et al*., 2003).

References

Fellbaum, C. (ed.). 1998. *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press.

Guthrie, L., E. Walker, and J. Guthrie. 1992. Document Classification by Machine: Theory and Practice. *Memoranda in Cognitive and Computer Science,* MCCS-92-235. Las Cruces, NM: Computing Research Laboratory, New Mexico State University

Dekang Lin, Shaojun Zhao, Lijuan Qin and Ming Zhou. 2003. Identifying Synonyms among Distributionally Similar Words. In *Proceedings of IJCAI-03*, pp.1492-1493.

Mahesh, K., and S. Nirenburg. 1995. A Situated Ontology for NLP. *Proceedings of the Workshop on Basic Ontological Issues in Knowledge Sharing*, International Joint Conference on Artificial Intelligence (IJCAI-95), Montreal, Canada. August, 1995.

Manning, Christopher, and Henrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*. Cambridge, MA: MIT Press.

Waikato Environment for Knowledge Analysis. 1999–2004. University of Waikato, New Zealand. (http://www.cs.waikato.ac.nz/ml/weka)

Weischedel, R., M. Meteer, R. Schwartz, L. Ramshaw, and J. Palmucci. 1993. Coping with Ambiguity and Unknown Words through Probabilistic Models. *Computational Linguistics*, 19(2): 359-382.